OIPE JC16
JAN 2 2 2002
PATENT & TRADEMARK OFFICE

Document #: 704490

M&C Folio: USP82149A

AUTOMATIC TEXT CLASSIFICATION SYSTEM

The present invention relates to an automatic text classification system, and more specifically to a system for automatically classifying texts in terms of each of a plurality of qualities in a manner such that the classified texts can be automatically retrieved based on a specified one or more of the plurality of qualities. The invention also relates to a retrieval system using the plurality of qualities.

A variety of methods are known for automatically classifying and/or analysing text, including keyword searching, collaborative filtering, and natural language parsing.

Keyword searching methods operate by simply looking for one or more keywords in a text and then classifying the text based on the occurrence (or non-occurrence) of the keywords. Keyword searching methods, however, suffer from the drawbacks that the main concept of a given text may be unrelated to the keywords being searched, and/or that a particularly relevant text may not contain the keywords being searched.

Collaborative filtering methods work by attempting to make recommendations and/or classifications based on matching overlapping results. For example, if a collaborative filtering system were used to analyse a series of questionnaires asking people to name their favourite musicians, the system would analyse the questionnaires by looking for an overlap in one or more of the musicians named in respective questionnaires. If an overlap were found between two questionnaires, the other musicians named by the author of the first questionnaire would be recommended to the author of the second questionnaire, and vice versa. The drawback of collaborative filtering, however, is that it assumes that people's tastes that are similar in one respect are also similar in other respects. That is, collaborative filtering methods fail to take into account the underlying qualities that define people's tastes.

The present application is a continuation-in-part of application serial number 09/615,295 filed July 13, 2000, and claims the benefit of said application serial number 09/615,295.

a modifying word is a common word that adds meaning to a main stem word; and

each word stem sequence comprises a main stem word and one or more previous words that are either modifying words or other main stem words.

79. The retrieval system according to claim 75, further comprising a graphical user interface for enabling input of the user preference data.

80. A system for producing training data comprising:

means for identifying lexical units and lexical unit sequences from each of a plurality of training texts that have been pre-classified with respect to each of a plurality of qualities; and

means for calculating a distribution value of each identified lexical unit and lexical unit sequence in each training text with respect to each of the plurality of qualities.

81. A method of producing training data comprising:

identifying lexical units and lexical unit sequences from each of a plurality of training texts that have been pre-classified with respect to each of a plurality of qualities; and

calculating a distribution value of each identified lexical unit and lexical unit sequence in each training text with respect to each of the plurality of qualities

82. A carrier medium carrying computer readable code for controlling a processor to carry out the method of any one of claims 14 to 26, and 48 to 51, ~~54, 55 or 57.~~

~~83. A carrier medium carrying computer readable code for controlling a computer to function as the system as claimed in any one of the claims 58 to 79.~~